data mining, medical data processing, rule sets

Ewa SZPUNAR-HUK*

PROBLEMS OF MEDICAL DATA MINING

The article discusses the main problems connected to the specificity of medical aspects, especially as concerns the quality and means of selection of data and tools used for constructing classification systems. Special attention is devoted to the risks inherent in direct application of classical knowledge extraction algorithms (such as the algorithms for constructing decision trees) to medical data. The article describes some attempts at solving emerging problems and points to the need for analysis of classifiers with regard to more than just their potential redundancy and mutual exclusion. The article also proposes two functions, useful for analysing rule sets with focus on data semantics.

1. THE ISSUES OF MEDICAL DATA MINING

The Knowledge Data Discovery (KDD) domain has been under development for over 10 years. During that time many universal algorithms for locating dependencies in information sets were created. So far, however, there has been little focus on dedicated tools. Medical applications pose highly specific demands: knowledge extraction algorithms in this area should provide for the diversity of types, quality and semantics of data being processed. Current techniques fail to fully live up to such conditions. As examples we can point to popular classification algorithms (CART, C4.5) and the decision tree pruning methods which attach more weight to generalization than to accuracy. There is a need for better data mining algorithms which would suit medical data processing tasks. At the same time, it becomes imperative to assess the correctness of current systems and methods employed in their creation. Medical data mining methods should focus on patient welfare and safety rather than on performance issues.

1.1. DATA-RELATED PROBLEMS

Many of the problems involved in applying current DM algorithms in medical areas result from the specificity and quality of medical data. Analyses usually only consider basic facts, which constitute a small fraction of the total set of information describing a particular patient. This is the primary reason behind the unreliability of computerized diagnoses. It seems that limiting the error ratio requires that more and more patient information be gathered and processed – such as environmental characteristics, lifestyle, eating habits and

* Department of Computer Science, Wroclaw University of Technology, ewa.szpunar@.*pwr.wroc.pl*.

other elements which may impact that person's clinical state. At the same time, we should also consider disease history and past treatments, which may significantly influence numerous symptoms.

Such a wide, multidimensional analysis of the patient's state is, however, currently impossible to conduct due to the emergence of other problems related to the accessibility and incompleteness of data. Analyses show that most records drawn from medical databases do not contain all the useful information, because not all patients are subject to the same types of examinations. This can result from financial constraints and from the fact that each doctor may order a different set of examinations to be administered, based on his/her personal experience. There are, of course, many ways of getting around the problem of missing data, but none of the accepted practices seems suitable for medical data processing. We cannot completely reject incomplete records, because we would end up discarding a significant amount of useful data; at the same time we cannot pick missing values at random, since this would endanger the patient. Duplicating records and filling the empty fields with all possible values of relevant attributes does not appear to be a suitable course of action either, as it would introduce false dependencies, not occurring in real life.

Additional problems result from the heterogeneity of medical data, as well as from ethical and legal aspects (described on [3]).

1.2. TOOL INADEQUACY

The analysis of medical data is typically aimed at constructing classification mechanisms, often expressed as decision trees or rule sets, to support qualified personnel in the decision making process. Unfortunately, due to the inadequacy of current knowledge extraction tools for solving medical problems, their results may prove dangerous to patients. It is therefore imperative to verify automated diagnoses as well as the chains of reasoning which ultimately lead to particular conclusions.

As an example, let us consider the oversimplification of decision trees by leading data classification algorithms (C4.5, CART). These algorithms frequently form the basis for knowledge bases in applied expert systems. Decision tree pruning in such cases relies on two techniques: Occam's razor and the removal of nodes not substantiated by an appropriate amounts of data. The former technique eliminates overly convoluted hypotheses, which may result in discarding leads that would otherwise help correctly diagnose complex cases. The latter method rejects dependencies which are not substantiated to a sufficient degree by actual data; however at the risk of omitting common cases that are nonetheless inadequately represented in test data sets (i.e. due to mistakes in data gathering approaches).

What is worse, current classification methods tend to neglect the issue of data semantics? Recent attempts at including such semantics in diagnostic systems rely mostly on attaching relevancy factors to individual attributes and decision classes, so that, for instance, when several conditions are diagnosed simultaneously, the gravest of them may be singled out for intensive treatment. Unfortunately, such methods impact the reasoning process itself, by neglecting the consequences of current and past decisions.

2. LOOKING FOR SOLUTIONS

In spite of the growing focus of researchers on issues pertinent to data mining in medical applications, many problems remain unsolved. There is an ongoing search for proper comparison and verification techniques that might be applied to the results delivered by diagnostic systems as well as to the methods used in their development. In addition, new algorithms and heuristics for extracting knowledge from medical databases are considered. Mostly, however, the discussions concentrate on problems, which stem from the imperfections and multidimensional nature of medical data, as well as on the ever-important issue of semantics in medical data sets.

2.1. SOLVING DATA-RELATED PROBLEMS

Most pertinent works tend to focus on medical data heterogeneity. Simultaneous consideration of multiple data sources for the purposes of reasoning requires the development of highly complex data translation and unification methods – hence the attempts at developing and promoting standardization in medical data storage and processing [3]. The heterogeneity of medical data is, however, also expressed in the selection of semantics for individual data sets. Gauging and capitalizing on the importance of selected data pieces, such as X-ray and EMR images, ECG readouts and disease histories for the purposes of augmenting the diagnostic process proves a difficult task – one that should also consider the circumstances and periods in which information is gathered [11]. All these problems call for new methods of analyzing multidimensional data.

In parallel, we consider the issue of incompleteness and inconsistency of medical databases. Due to the inability of omitting incomplete or inconsistent data (since this problem plagues a significant percentage of all medical data used for classification), there is a need for effective methods, which would enable their use. Given the lack of a basis for application of typical methods of filling in missing values (applied in typical data mining aspects), we research new techniques, based on the principle of incremental knowledge extraction. This is particularly important when considering the potential threat to human life as a result of including incorrect data in knowledge bases. At the same time, attempts are made at characterising the noise types in medical databases, preparing ground for their effective elimination [7,8].

2.2. SOLVING TOOL-RELATED PROBLEMS

Despite the fundamental nature of the works described above, the mainstream of medical data mining research remains closely tied with the assessment of existing classifiers and developing new methods of their creation. Gamberger’s works from the late nineties, combined with the unexpectedly accurate results achieved by complex classifiers when compared with “canonical” machine learning algorithms (C4.5, CART) have made researchers aware of the fact that radical decision tree pruning methods cannot be used without regard for their influence on the level of user safety in the resultant knowledge

bases [6,7]. Hence, some criteria governing the usage of MDL (Minimum Description Length) techniques have been formulated, basing on the Ockham's Razor principle. This shadow, cast on standard data mining, has resulted in the emergence of new techniques, such as grafting [17], and has led to an intensification of research in the area of heterogeneous classification systems, unifying various applied AI methods (including neural networks, genetic programming and fuzzy logic). More attention is also being devoted to analyzing the influence of proposed methods on their long-run properties (highly important from the patients' point of view).

The perceived necessity of taking into account the dynamic character of knowledge derived from medical information sets has spawned significant interest in active learning and reasoning methods - aware of temporal dependencies among pieces of data. This approach also enables us to supplement medical decision support systems with modules that assess a number of phenomena closely tied to long-term changes in the processes and events inherent in patients' lives. This ability, in turn, proves important for monitoring the creeping evolution of medical conditions in patients, enabling a division of cases initially treated as noise into distinct classes, subsequently analyzed in the reasoning process. Such system behaviour may be useful for long-term (even continuous) environmental monitoring, but also for extending the set of available diagnostics techniques [9, 14].

2.3. RESEARCH OF EXISTING REASONING SYSTEMS

Attempts at constructing and applying decision support systems have been undertaken since the early sixties. Yet, the emergence of new problems and threats linked to widespread application of automated decision support methods calls into question their continued usefulness. We should therefore resort to multifaceted assessment of existing rule-based knowledge bases and decision trees. Effort is currently being invested in various types of gradation of rule sets - i.e. based on their similarity to other rules or the degree of evenness and accuracy of decision space fragmentation. Some researchers even try to think "out of the box" by focusing on the analysis of rule relevancy and interestingness for future knowledge users [5, 10].

Nevertheless, experience shows that the analysis of knowledge bases (in particular of medical ones) cannot proceed without prior verification of their semantics. Unfortunately, due to the large sizes of applied reasoning systems, any human expert-based analysis of their rules would be difficult at best. Yet, despite the difficulties, this area is rife with interesting solutions - as an example, we can point to some grouping methods [1,4] or to a new method of transparent rendering of large rule sets through specific grouping strategies, developed at the University of Silesia [15]. It makes available to significantly increase performance of KDD process and knowledge management.

3. RULE SET ANALYSIS

The above issues are reflected in ongoing research on specific analysis methods for knowledge bases. The author's research has so far resulted in proposing two methods for assessing rule set similarity, based on the properties of decision spaces which they define within the attribute value spectrum. The first such method, called the rule set density function, can prove helpful in locating those areas of the rule set decision space which are quantified with insufficient precision. Another method, the rule set distance function, is meant to augment approximate comparisons of rule sets basing on their semantics.

Both proposals are part of the wide subject of seeking selective sampling techniques for data space exploration, required for proper functioning of active learning methods [16]. They seem indispensable for gathering knowledge through minimizing the data sets required for operation of classifiers. Preliminary research points to the conclusion that, contrary to direct selection methods described in literature (e.g. uncertainty sampling [2], confidence measures [16] and weighted sampling [13]), the described methods allow for natural utilization of the knowledge on training data as well as of the semantics of classifiers which describe that data.

3.1. THE RULE SET DENSITY FUNCTION

The first of two proposed functions - the density function - relies on determining the number of rules covering a selected fragment of the data domain. It can be expressed in the following way:

$$D = \frac{N}{V} \qquad (1)$$

where:

D – the density of the selected domain fragment

N – the number of rules covering the selected data domain area

V – the size of the selected area for a given volume unit

Figure 1 shows a sample data domain with two attributes, each assuming values between 0 and 10. Data areas covered by individual rules are represented as numbered rectangles (the number indicates the rule). In this sample, the density function (given a volume unit of 1) is equal to 1/8 for area A_1 and 1/4 for area A_2.

Thus defined, the density function allows us to determine which data domain areas are probably described by a small number of rules and therefore likely underrepresented in training data sets or neglected by the expert (if the rules are supplied by a human agent). On the other hand, areas with a large number of adjacent rules might be afflicted by severe noise or by classifiers having adapted too much to training data at the cost of versatility and ability to generalize (such classifiers will typically prove unable to correctly interpret data from outside of the training sets).

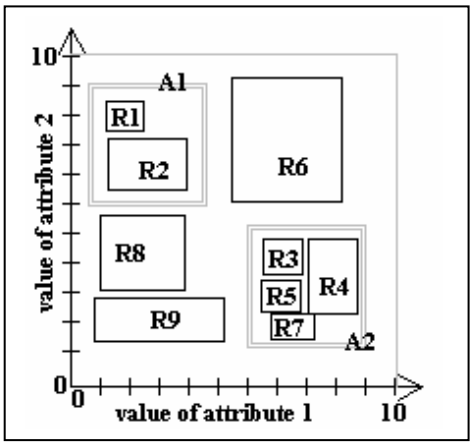

Fig. 1. Sample distribution of rule decision spaces in the data domain.

3.2. THE RULE SET DISTANCE FUNCTION

Another proposed function is the rule set distance function. It allows us to gauge the respective positioning of decision spaces corresponding to two selected rule sets. This function relies on the estimation of a Euclidean distance between selected examples drawn from rules decision spaces and treated as vectors. The distance function is heuristic in nature if the points are selected at random. The distance between rules R_1 and R_2 is thus expressed as:

$$L(R_1, R_2) = \min \left\{ L^*(R_1, R_2),\ L^*(R_2, R_1) \right\} \qquad (2)$$

where:

$$L^*(R_1, R_2) = \frac{1}{N} \sum_{i=1}^{N} \min^* (p_i, R_2),\ p_i \in T(R_1) \qquad (3)$$

is the distance between the decision space of rule R_1 and that of rule R_2, $T(R_1)$ is a set of N test samples p belonging to the decision space of rule R_1, while min* is the smallest Euclidean distance between some point p_i belonging to the decision space of rule R_1 and the selected test points covered by rule R_2.

Figure 2 presents a sample distribution of the decision spaces of eight rules in the data domain (for two attribute values, each between 0 and 80). Table 1 contains a matrix of distance function values for given pairs of rules. The analysis of such a matrix allows us to determine (in a manner similar to the density function case) which areas of the data domain are densely populated by rules: such areas typically comprise a set of rules placed close to each other in a “clump” (in Figure 2 this is true for the area covered by rules R_5, R_6, R_7 and R_8). Densely populated areas are often difficult to classify and typically involve significant levels of noise.

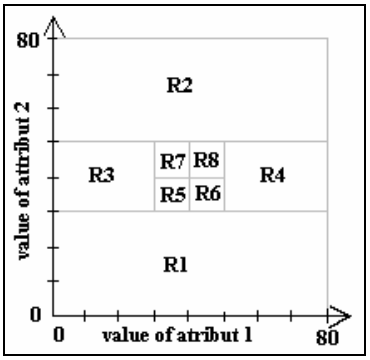

Fig. 2. Sample distribution of rule decision spaces in the data domain.

Table 1. Distance function values for pairs of rules.

Rule	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**
R1	0.38	34.66	25.98	25.65	24.33	21.37	31.63	31.62
R2	34.66	0.39	25.21	26.13	31.98	29.63	23.42	24.42
R3	25.98	25.21	0.19	35.21	15.64	15.55	16.02	24.95
R4	25.65	26.13	35.21	0.19	25.49	15.73	25.22	15.95
R5	24.33	31.98	15.64	25.49	**0.08**	**0.09**	**4.99**	**7.72**
R6	21.37	29.63	15.55	15.73	**0.09**	**0.12**	**6.36**	**6.37**
R7	31.63	23.42	16.02	25.22	**4.99**	**6.36**	**0.08**	**5.05**
R8	31.62	24.42	24.95	15.95	**7.72**	**6.37**	**5.05**	**0.08**

The distance function may additionally be applied to comparing two different sets of rules, i.e. when one of them comes from an expert while the other is determined by data mining methods. The greater the distances between rules from different sets, the greater the differences between data domain areas covered by these rules.

Another interesting experiment involving the comparison of two rule sets (i.e. for evaluation of temporal changes in data) may be based on dividing these sets into subsets for individual classes of objects, which are then subject to assessment. If the distance function values for two subsets from the same class are significant, then we can effectively conclude that a major shift in the knowledge relating to such data has taken place.

Additionally, the distance function allows us to locate rules of interest - for example those, for which decision spaces are situated far away from spaces corresponding to other rules, i.e. because they refer to highly peculiar conditions or unusual disease symptoms.

4. CONCLUSION

The presented discussion on knowledge extraction from medical databases is merely a short summary of the ongoing efforts in this area. It does, however, point to interesting directions of research, where the aim is to create data mining tools well suited to the crucial demands of medical diagnostic systems (patient safety in particular). This summary serves

as the backdrop for presenting the author's own research efforts as an attempt at contributing to the development of automatic diagnostic systems.

BIBLIOGRAPHY

[1] BING LIU, MINQUING HU, WYNNE HSU, Multi-Level Organization and Summarization of the Discovered Rules, KDD-2000, 2000

[2] CATLETT J., LEWIS D., Heterogeneous uncertainty sampling for supervised learning. In 11th International Conference on Machine Learning, 1994.

[3] CIOS K.J., MOORE G.W., Uniqueness of medical data mining, Artificial Intelligence in Medicine, Vol. 26, No 1-2,pp.1-24, September-October 2002

[4] DREWRY D. T., LIN GU, HOCKING A. B., KYOUNG-DON KANG, PFALTZ J. L., SCHUTT R. C., TAYLOR C. M., Current state of data mining, University of Virginia, Department of Computer Science, Technical Report,2002

[5] FREITAS A.A.: On rule interestingness measures. Knowledge-Based Systems journal 12 (5-6), 309-315. Oct. 1999.

[6] GAMBERGER D., LAVRAC N., Conditions for Occam's razor applicability and noise elimination. Proc. 9th European Conference on Machine Learning , pp. 108-123, Springer, Berlin,19997

[7] GAMBERGER D., LAVRAC N., KRSTATIC G.,SMUC T., Inconsistency tests for patients records in a coronary heart disease database, Proc. of Int. Symp. on Medical Data Analysis, pp. 183-189,2000

[8] GAMBERGER D., LAVRAC N., GROSELI C., Experiments with noise filtering in a medical domain. Proc. International Conference on Machine Learning, pp. 143-151,1999

[9] HARRIES M. B., SAMMUT C., HORN K., Extracting hidden context. Machine Learning, Vol 32,pp. 101-126, 1998

[10] HILDERMAN R, HAMILTON H., BARBER B., Ranking the interestingness of summaries from data mining systems, FLAIRS'99, Orlando, FL, May 1999

[11] KUKAR M., KONONENKO I., GROSELJ C., KRALJ K., FETTICH J., Analysing and improving the diagnosis of ischaemic heart disease with machine learning, Artificial Intelligence in Medicine, Vol 16,pp. 25-50, 1999

[12] LINDENBAUM M., MARKOVITCH S., RUSAKOV D., Selective Sampling for Nearest Neighbour Classifiers, Machine Learning,Volume 54, 2004

[13] LIU H., MOTODA H., YU L. *Feature selection with selective sampling*. In Proceedings of the Nineteenth International Conference on Machine Learning, pages 395 - 402, 2002

[14] SHAHAR Y., CHENG C., Knowledge-Based Visualization of Time-Oriented Clinical Data. Proceedings of the 1998 AMIA Fall Symposium, Orlando, FL1998

[15] SIMINSKI R., WAKULICZ-DEJA A., Lokalna i globalna weryfikacja regulowych baz wiedzy w oparciu o koncepcje jednostek decyzyjnych, Mater. III Konf. MSK, Kraków, 2001

[16] THRUN S., Exploration in active learning. In M. A. Arbib, editor, The Handbook of Brain Theory and Neural Networks, pages 381-384. MIT Press, Cambridge, MA, 1995.

[17] WEBB G. I., Decision tree grafting, Fifteenth International Joint Conference on Artificial Intelligence, pp. 846--851, Morgan Kaufmann, Japan, 1997.